jh240058

Study on the Real Effect of Non-Blocking Collective Communications

Takeshi Nanri (Kyushu U.), Kengo Nakajima (U. Tokyo), Richard Vuduc (Georgia Tech.), Takeshi Fukaya (Hokkaido U.), Hiroyuki Takizawa (Tohoku U.), Osamu Tatebe (U. Tsukuba), Daisuke Takahashi (U. Tsukuba), Toshihiro Hanawa (U. Tokyo), Shinji Sumimoto (U. Tokyo), Maddegedara Lalith (U. Tokyo), Rio Yokota (Tokyo Tech.), Takahiro Katagiri (Nagoya U.), Keiichiro Fukazawa (Kyoto U.), Susumu Date (Osaka U.), Takashi Soga (Osaka U.), Yoshiyuki Morie (Teikyo U.), Richard Graham (NVIDIA), Martin Schulz (TUM), Bengisu Elis (TU Munich), Dennis Herr (TUM), Hari Subramoni (Ohio State U.), Aamir Shafi (Ohio State U.), Kaushik Kandadi suresh (Ohio State U.), Nathaniel Shineman (Ohio State U.), Benjamin Michalowicz (Ohio State U.), Tu Tran (Ohio State U.), Shulei Xu (Ohio State U.), Bharath Ramesh (Ohio State U.), Felix Wolf (TU Darmstadt), Gerhard Wellein (NHR), Gerardo Cisneros-Stoianowski (NVIDIA), Brody Williams (NVIDIA), Yong Qin (NVIDIA), Fabian Czappa (TU Darmstadt), Ayesha Afzal (NHR), Takeo Narumi (Kyushu U.)

Motivation

- Collective communication is the significant causes of scalability degradation in HPC.
- NBC (Non-Blocking Collective communication) is expected to be a means to overlap this collective communication with computation and hide the communication time, but its use is currently limited to a small number of applications.
- This project provides programmers with correct knowledge about the usage and performance characteristics of NBC and the effect of communication hiding in real applications.

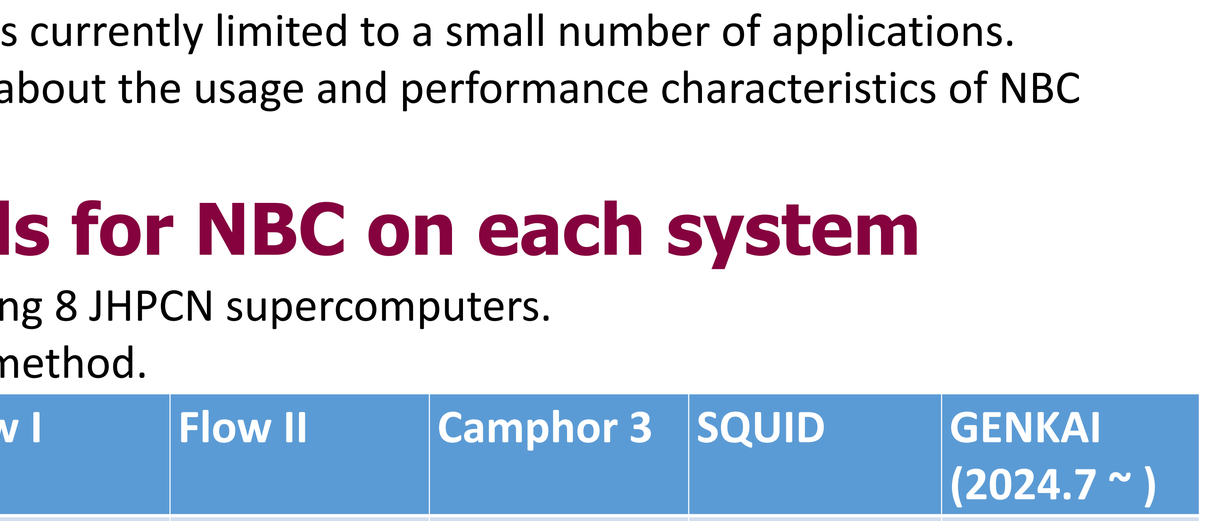
Topic 1: Available progress methods for NBC on each system

- 4 progress methods for NBC will be examined on the following 8 JHPCN supercomputers.
- Documents of the usage will be released for each available method.

Method	AOBA-S	Wisteria-Oddysey	TSUBAME 4.0	Flow I	Flow II	Camphor 3	SQUID	GENKAI (2024.7 ~)
SHARP	in study	NG	OK	NG	NG	in study	in study	
Tofu Barrier	NG	in study	NG	in study	NG	NG	NG	
Assistant core	NG	OK	NG	OK	NG	NG	NG	
Progress thread	OK	OK	OK	OK	OK	OK	OK	

Topic 2: Trends of the effect of overlapping by NBC

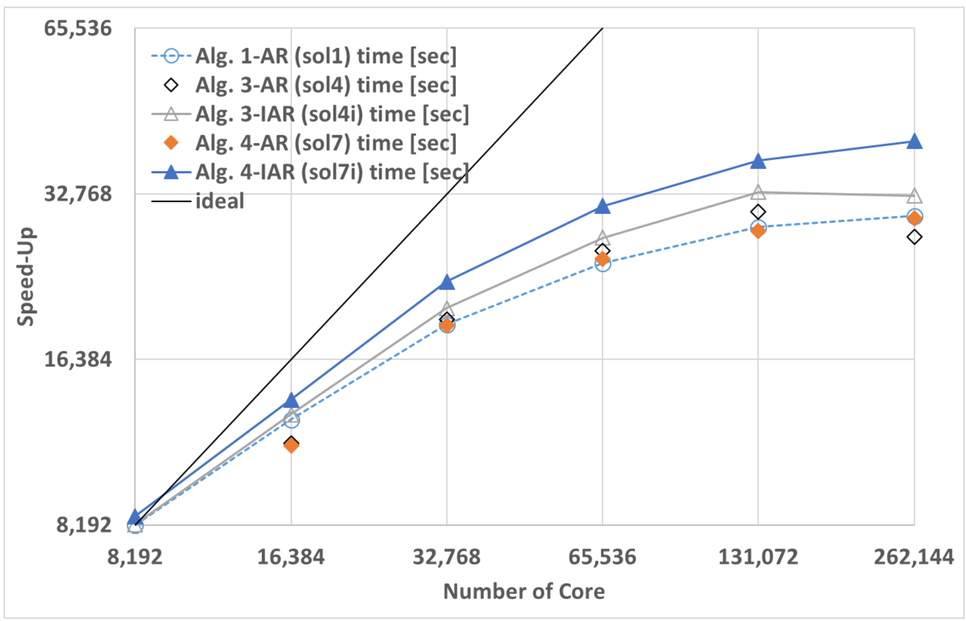
- Effect of overlapping by NBC depends on various conditions such as interconnect, progress method, PPN (number of processes per node), etc.
- Preliminary results
  - Overlapping ratio on TSUBAME 4.0 (OSU Micro-benchmarks, MPI_Iallreduce)
  - Other available methods will be examined on each system
- In addition to the overlap ratio, the effects on the total execution time will be investigated with another benchmark.



Topic 3: Investigation of communication hiding algorithms with NBC

- Description

Krylov subspace iterative methods, like Conjugate Gradient (CG), are essential in scientific computing for solving large-scale linear equations with sparse matrices from FEM, FDM, and FVM. Efficient execution and minimized communication are crucial for large-scale parallel computing. Research on Communication-Computation Overlapping (CC-Overlapping) has led to pipelined methods that reduce communication overhead by overlapping communications and computations. Pipelined CG changes the sequence of Krylov iterations using recurrence relations, allowing collective communication for dot products to overlap with heavier computations. Implemented on Intel Xeon Phi systems, pipelined CG achieved a 40% speed-up over original CG. This project evaluates the performance on NVIDIA SHARP systems and aims to develop stable algorithms for lower precision computing. Future plans include developing GPU versions and optimizing collective communications for applications like FFT.



Alg.1 Original PCG
Alg.3 Pipelined CG
Alg.4 Gropp's CG

IAR: MPI_Iallreduce
AR : MPI_Allreduce

Performance of Preconditioned CG Solvers for Parallel FEM Application with Pipelined Algorithms on Oakforest-PACS System at JCAHPC using Intel Xeon/Phi [Horikoshi et al. 2022]

- Schedule

(FY.2024) Evaluations on supercomputers with SHARP using CPU, Pipelined algorithms with mixed/lower precision
(FY.2025) GPU version of the codes, Evaluation/improvement of pipelined algorithm with mixed/lower precision
(FY.2026) Evaluations on supercomputers with SHARP using CPU/GPU